Classifier Ensemble Design for Imbalanced Data Classification: A Hybrid Approach

Uma R. Salunkhe*, Suresh N. Mali

*Department of Computer Engineering, Smt. Kashibai Navale College of Engineering, Savitribai Phule Pune University, Pune, India
bSinhgad Institute of Technology and Science, Pune, India

Abstract

Imbalanced learning for classification problems is the active area of research in machine learning. Many classification systems like image retrieval and credit scoring systems have imbalanced distribution of training data sets which causes performance degradation of the classifier. Re-sampling of imbalanced data is commonly used to handle imbalanced distribution as it is independent of the classifier being used. But sometimes they can remove necessary data of the class or can cause over-fitting. Classifier Ensembles have recently achieved more attention as effective technique to handle skewed data.

The focus of the work is to gain advantages of both data level and classifier ensemble approach in order to improve the classification performance. We present a novel approach that initially applies pre-processing to the imbalanced dataset in order to reduce the imbalance between the classes. The pre-processed data is provided as training dataset to the classifier ensemble that introduces diversity by using different training datasets as well as different classifier models. The experimentation conducted on the eight imbalanced datasets from KEEL repository helps to prove the significance of the proposed method. A comparative analysis shows the performance improvement in terms of Area under ROC Curve (AUC).

Keywords: Imbalanced data; re-sampling; classifier ensemble; under-sampling

1. Introduction

Classification is one of the important tasks in machine learning that is used to categorize the data. Many real world classification systems like fraud detection and credit scoring system face the critical problem of Imbalanced dataset where one class consist of very high number of instances relative to the other class. Hence imbalanced learning is gaining a great deal of attention of researchers in the domain of machine learning. Existing studies in imbalanced learning can be classified into four groups:

1877-0509 © 2016 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
Peer-review under responsibility of the Organizing Committee of CMS 2016
doi:10.1016/j.procs.2016.05.259
1. Data level approaches
2. Algorithm level approaches
3. Cost sensitive learning approaches
4. Classifier Ensemble techniques

First category aims to rebalance the distribution of classes by applying re-sampling techniques like under-sampling, over-sampling or hybrid approach. Recent studies using data level approaches have proved to improve the classification results but can raise some problems like over-fitting, removal of necessary data of the class etc. Algorithm level approaches modify the existing algorithms in order to handle the imbalanced data. The approach is beneficial if the learning algorithm is chosen intelligently. Third category combines the data and algorithm level approaches to gain benefits of both. Recent experimental studies show that classifier ensemble may improve the classification performance if we combine multiple diverse classifiers which disagree with each other.

Traditional ensemble methods work efficiently when provided with input training data set that has balanced distribution. Nowadays, many applications have imbalanced distribution of data i.e. different classes with unequal amount of data. This can result in the classifier model that is biased towards the majority class which has high number of instances than the other class. This raises need of some modifications in the existing ensemble techniques so that they can handle the imbalanced data in an efficient manner.

In this paper, we aim to improve classification performance by combining data level approach with proposed classifier ensemble technique. Data level approach involves re-sampling using over-sampling, under-sampling or both. We have used hybrid re-sampling technique in order to reduce the ‘Imbalance ratio’ which is defined as ratio of number of majority class instances to that of minority class. The re-sampled data is given for training the classifier ensemble model that is formed by using different classifier models and different training data subsets.

The rest of this paper is organized as follows. Section 2 reviews related work that describes the state-of-the-art ensemble techniques to handle imbalanced data sets. In Section 3, we introduce different classifier ensemble learning algorithms relevant to the study. Section 4 discusses the hybrid approach of classifier ensemble design. In Section 5, we present the details of experimental framework, that is, the algorithms that are used, the data-sets etc. In section 6, experimental analysis and the statistical tests that are used for the experimental study are discussed. Finally, section 7 concludes the paper.

2. Related Work

This section briefs the work carried out in the area of imbalanced data set classification by well known researchers in recent years.

Yubin Park et. al.\(^2\) presented two types of decision tree ensembles to handle issues faced by imbalanced data set. They have used an ensemble framework of $\alpha$ tree which gives higher AUROC values over different skewed data sets. A new splitting criterion has been introduced which uses diversification factor alpha ($\alpha$). There is a need to focus on the diversity of base classifiers.

Mikel Galar et. al.\(^3\) proposed a novel ensemble construction technique known as EUSBoost that is based on RUBOost technique and makes combined use of evolutionary under-sampling with boosting. Evolutionary under-sampling initially applies random under-sampling technique to the skewed data sets, which are evolved until there are no further improvements in currently best under-sampled data set. Use of evolutionary under-sampling method has shown improvement in the performance of base classifier.

Putthiporn Thanathamthee et. al \(^4\) introduced the concept of bootstrap re-sampling and AdaBoost technique to generate synthetic boundary data. The approach expands distribution of training data space in order to handle unseen
incoming data in efficient manner. Evaluation of proposed method in terms of overall accuracy, G-mean and F-measure shows the significant performance improvement compared to other methods.

Peng Cao et al.\(^5\) proposed novel method that combined hybrid probabilistic sampling with diverse random subspace ensemble. The instances with higher probability are most frequent in the class, and hence their removal may not result in loss of information. The instances to be under-sampled are selected according to the Gaussian distribution.

Xu-Ying Liu et al.\(^6\) introduced two algorithms known as EasyEnsemble and BalanceCascade which are more robust than many other class-imbalance learning methods. Evaluation in terms of Area under the ROC Curve (AUC), F-measure, and G-mean values shows performance improvement relative to many existing algorithms. But issues like lack of comprehensibility should be handled.

Gregory Ditzler et al.\(^7\) presented two ensemble based approaches known as Learn++.NIE and Learn++.CDS, for classification of imbalanced data in non stationary environment. Learn++.CDS uses Synthetic Minority Oversampling Technique (SMOTE) to rebalance the class distribution. Learn++.NIE generates ensemble of ensembles and hence is computationally expensive algorithm.

### 3. Classifier Ensemble

In medical applications, more opinions are taken from different doctors before making the final decision of any critical case. This helps to give the reliable result as decisions from multiple experts are combined in some way. This concept of combining decisions of multiple experts can also be applied to the classification of data and is known as ‘Classifier Ensemble (CE)’ or ‘Multiple Classifier System (MCS)’. Classifier Ensemble is combination of different individual classifiers in order to perform the classification task jointly. If those individual classifiers are diverse i.e. disagree with each other, then their random errors will cancel each other and will help to output correct decisions. Data to be classified is given as input to the number of individual classifiers to get number of predictions as output. The output predictions are then combined by using different methods such as voting, average, weighted voting etc.

Classifier ensemble can be formed by using different techniques which can be categorized into following categories:\(^8\):

1. Using different training sets: Introduce diversity by partitioning training dataset into N subsets and training individual classifier with different subsets.
2. Using different feature subsets: Introduce diversity by training individual classifier with different subset of features.
3. Using different classifier models: Introduce diversity by combining different individual classifier.
4. Using different combination schemes: Introduce diversity by using different combination schemes.

Following subsections give a brief introduction to the ensemble approaches that are relevant to this study.

#### 3.1 Bagging (Bootstrap Aggregating)

Bagging is a classifier ensemble that is formed by using different training data subsets. It is a combination of bootstrapping and averaging in which multiple versions of predictor are generated and combined to finally generate aggregated predictor. Individual base classifiers are independently trained on different training sets known as bootstrap samples. Bootstrap samples are formed by randomly picking some samples with replacement from the original samples of the training set. Hence, bagging is preferred to use with unstable algorithms, where the small changes in the training set, result in large changes in the output of that system\(^9\).

#### 3.2 Stacking

Stacking is the classifier ensemble that is formed by using different classifier models. It combines multiple classifiers generated by different learning algorithms L1,…..Ln. Those learning algorithms are trained on a single
dataset and a set of base level classifiers C1,…..Cn is generated. The predictions of the individual classifiers are combined by Meta level classifier. This method reduces the risk of over-fitting.

4. Hybrid Approach for Classifier Ensemble Design

The proposed work is carried out in two phases:

1. Re-sampling of the imbalanced data
2. Classifier Ensemble formation

Steps of the hybrid approach

1. Re-sample the imbalanced data
   a. Over-sampling using Synthetic Minority Over-sampling Technique (SMOTE)
      SMOTE\(^9\), a well known over-sampling technique is applied to the imbalanced data set in order to increase the number of samples of minority class. This will help to reduce the imbalance ratio of the data set.
   b. Under-sampling using modified random under-sampling technique
      Random under-sampling technique randomly selects and deletes some instances of majority class in order to decrease the imbalance ratio of the data set. But this may remove some necessary instances of the data set. To overcome this limitation, we have modified the algorithm by initially identifying the necessary data of the majority class and then applying random under-sampling to the remaining data.

2. Classifier Ensemble formation
   a. Using different training set
      Bagging classifier ensemble is constructed using different training data subsets known as bootstrap samples. J48 is used as base classifier and voting method is used to combine the predictions of individual classifiers.
   b. Using different Classifier Model
      StackingC classifier ensemble is formed by using different classifier model. In our work, we have used three classifier models as base classifiers namely J48, LogisticRegression and Bagging. Here bagging works as individual classifier but bagging itself is a classifier ensemble that has been formed by using different training subsets. Thus a hybrid approach is used to form the classifier ensemble.

5. Experimental Design

5.1 Experimental Setup

The experiments were carried out using Weka environment with its default parameters. Weka is an open source\(^{11}\) toolkit that provides a set of machine learning & pre-processing algorithms.

In this study we implemented hybrid data level approach for pre-processing the imbalanced data. This technique combines commonly used oversampling technique SMOTE with modified under-sampling approach. Preprocessed data is provided as input to ensemble of classifiers that is constructed by using different training datasets as well as different classifier models. StackingC ensemble is constructed by using different classifier models namely J48, LogisticRegression and Bagging. All experiments were carried out using 5 fold cross validation. Experimental results of the proposed approach are compared with the results of the classifier ensemble that is constructed only by using different training sets.

5.2 Experimental Data sets
For experimentation, we have chosen eight imbalanced data sets that are publicly available in KEEL repository. We have chosen eight data sets with varying imbalance ratios ranging from 2 to 80. Details of those data sets are given in table 1. For each data-set, following details are specified:

- Number of examples
- Number of majority class examples
- Number of minority class examples
- Imbalance Ratio

<table>
<thead>
<tr>
<th>Data set</th>
<th># Examples</th>
<th># Majority</th>
<th># Minority</th>
<th>IR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abalone-20_vs_8-9-10</td>
<td>1916</td>
<td>1890</td>
<td>26</td>
<td>72.69</td>
</tr>
<tr>
<td>ecoli-0-1-4-6_vs_5</td>
<td>280</td>
<td>260</td>
<td>20</td>
<td>13</td>
</tr>
<tr>
<td>car-good</td>
<td>1728</td>
<td>1659</td>
<td>69</td>
<td>24.04</td>
</tr>
<tr>
<td>Iris0</td>
<td>150</td>
<td>100</td>
<td>50</td>
<td>2</td>
</tr>
<tr>
<td>Flare-f</td>
<td>1066</td>
<td>1023</td>
<td>43</td>
<td>23.79</td>
</tr>
<tr>
<td>Glass6</td>
<td>214</td>
<td>185</td>
<td>29</td>
<td>6.38</td>
</tr>
<tr>
<td>new-thyroid1</td>
<td>215</td>
<td>180</td>
<td>35</td>
<td>5.14</td>
</tr>
<tr>
<td>Pima</td>
<td>768</td>
<td>500</td>
<td>268</td>
<td>1.87</td>
</tr>
</tbody>
</table>

5.3 Evaluation Parameters

Classification systems can be evaluated by using standard parameters such as accuracy, G-mean, F-measure, type-I error, type-II error, AUC (Area under ROC curve) etc. These parameters can be derived by using confusion matrix. Table 2 presents a confusion matrix for a binary class.

<table>
<thead>
<tr>
<th>Predicted as positive</th>
<th>Predicted as negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive Class</td>
<td>True Positive</td>
</tr>
<tr>
<td>Negative Class</td>
<td>False Positive</td>
</tr>
</tbody>
</table>

Previous studies show that accuracy does not differentiate between the number of correctly classified instances of majority and minority classes. Hence it is not appropriate measure for imbalanced data set classification and Area under the ROC curve (AUC) has been suggested as an appropriate measure. In this paper, we have used AUC as evaluation parameter which is defined as arithmetic average of the mean predictions for each class\textsuperscript{12}. AUC can be represented as

\[
AUC = \frac{\text{Sensitivity} + \text{Specificity}}{2}
\]  
(1)

\[
\text{Sensitivity} = \frac{TP}{TP + FN}
\]
(2)
We have used Wilcoxon signed rank test to evaluate the significance of the proposed technique. The Wilcoxon signed rank test is a non-parametric statistical procedure for comparing two samples that are paired. This method is based on difference scores that not only considers signs of the differences but also takes into account the magnitude of the differences.

6. Results and Discussion

Table 3 shows the experimental results of classification algorithm when classifiers are constructed by using different techniques. Initially imbalanced data is re-sampled using hybrid re-sampling approach in order to reduce its imbalance ratio. The output dataset is used to train the classifier ensemble that is formed by using two different techniques i.e. using different training datasets and different classification models. StackingC ensemble is formed by using different classifier models namely J48, LogisticRegression and Bagging.

The experiments are carried out for different values of parameter N that indicates percentage of over-sampling to be done for SMOTE.

Model-1: ‘Bagging’ classifier ensemble with base classifier J48
Model-2: ‘StackingC’ classifier ensemble with base classifiers J48, LogisticRegression and Bagging

Table 3 Results of two classifier ensembles

<table>
<thead>
<tr>
<th>Data set</th>
<th>Classifier Ensemble</th>
<th>AUC 100</th>
<th>AUC 200</th>
<th>AUC 300</th>
<th>AUC 400</th>
<th>AUC 500</th>
</tr>
</thead>
<tbody>
<tr>
<td>abalone-20_vs_8-9-10</td>
<td>Model-1</td>
<td>0.905</td>
<td>0.956</td>
<td>0.971</td>
<td>0.966</td>
<td>0.967</td>
</tr>
<tr>
<td></td>
<td>Model-2</td>
<td>0.932</td>
<td>0.977</td>
<td>0.979</td>
<td>0.967</td>
<td>0.971</td>
</tr>
<tr>
<td>ecoli-0-1-4-6_vs_5</td>
<td>Model-1</td>
<td>0.94</td>
<td>0.982</td>
<td>0.998</td>
<td>0.994</td>
<td>0.992</td>
</tr>
<tr>
<td></td>
<td>Model-2</td>
<td>0.966</td>
<td>0.983</td>
<td>0.998</td>
<td>0.992</td>
<td>0.994</td>
</tr>
<tr>
<td>car-good</td>
<td>Model-1</td>
<td>0.967</td>
<td>0.985</td>
<td>0.982</td>
<td>0.986</td>
<td>0.982</td>
</tr>
<tr>
<td></td>
<td>Model-2</td>
<td>0.981</td>
<td>0.994</td>
<td>0.984</td>
<td>0.989</td>
<td>0.984</td>
</tr>
<tr>
<td>iris0</td>
<td>Model-1</td>
<td>0.995</td>
<td>0.99</td>
<td>0.995</td>
<td>0.995</td>
<td>0.995</td>
</tr>
<tr>
<td></td>
<td>Model-2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>flare-f</td>
<td>Model-1</td>
<td>0.879</td>
<td>0.803</td>
<td>0.846</td>
<td>0.807</td>
<td>0.855</td>
</tr>
<tr>
<td></td>
<td>Model-2</td>
<td>0.897</td>
<td>0.808</td>
<td>0.847</td>
<td>0.821</td>
<td>0.865</td>
</tr>
<tr>
<td>glass6</td>
<td>Model-1</td>
<td>0.953</td>
<td>0.977</td>
<td>0.991</td>
<td>0.979</td>
<td>0.989</td>
</tr>
<tr>
<td></td>
<td>Model-2</td>
<td>0.99</td>
<td>0.976</td>
<td>0.998</td>
<td>0.997</td>
<td>0.989</td>
</tr>
<tr>
<td>New-thyroid1</td>
<td>Model-1</td>
<td>0.974</td>
<td>0.998</td>
<td>1</td>
<td>0.987</td>
<td>0.997</td>
</tr>
<tr>
<td></td>
<td>Model-2</td>
<td>0.98</td>
<td>0.998</td>
<td>0.999</td>
<td>0.995</td>
<td>1</td>
</tr>
<tr>
<td>pima</td>
<td>Model-1</td>
<td>0.885</td>
<td>0.882</td>
<td>0.916</td>
<td>0.908</td>
<td>0.893</td>
</tr>
<tr>
<td></td>
<td>Model-2</td>
<td>0.892</td>
<td>0.885</td>
<td>0.918</td>
<td>0.912</td>
<td>0.907</td>
</tr>
</tbody>
</table>

Fig. 1 plots AUC against percentage of over-sampling (N) for two imbalanced data sets i.e. abalone-20_vs_8-9-10, ecoli-0-1-4-6_vs_5 respectively.
Fig 1. AUC performance on the two data sets at different percentage of oversampling (a) abalone-20_vs_8-9-10 (b) ecoli-0-1-4-6_vs_5

Analysis of the above graphs clearly shows that in almost all cases the Model-2 that uses hybrid classifier ensemble under-sampling technique results in the improvement of AUC. Although the improvement in AUC value is relatively smaller, it is beneficial for skewed data set where accuracy of minority class is extremely important.

We have evaluated the statistical significance of our performance results by using Wilcoxon signed rank test. The significant test was applied on the AUC performance measure and at the significant level 0.05. For eight datasets the critical value is 3 and hence value of T should be less than or equal to 3. Table 4 shows the example of the significant test of AUC performance between our method and Model-1. T can be expressed as

\[ T = \min \{R^+, R\} \]

Table 4 An example of significant test of AUC performance between Model-1 and Model-2

<table>
<thead>
<tr>
<th>Data set</th>
<th>Model-1</th>
<th>Model-2</th>
<th>Difference</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abalone-20_vs_8-9-10</td>
<td>0.905</td>
<td>0.932</td>
<td>0.027</td>
<td>7</td>
</tr>
<tr>
<td>ecoli-0-1-4-6_vs_5</td>
<td>0.94</td>
<td>0.966</td>
<td>0.026</td>
<td>6</td>
</tr>
<tr>
<td>car-good</td>
<td>0.967</td>
<td>0.981</td>
<td>0.014</td>
<td>4</td>
</tr>
<tr>
<td>Iris0</td>
<td>0.995</td>
<td>1</td>
<td>0.005</td>
<td>1</td>
</tr>
<tr>
<td>Flare-f</td>
<td>0.879</td>
<td>0.897</td>
<td>0.018</td>
<td>5</td>
</tr>
<tr>
<td>Glass6</td>
<td>0.953</td>
<td>0.99</td>
<td>0.037</td>
<td>8</td>
</tr>
<tr>
<td>new-thyroid1</td>
<td>0.974</td>
<td>0.98</td>
<td>0.006</td>
<td>2</td>
</tr>
<tr>
<td>Pima</td>
<td>0.885</td>
<td>0.892</td>
<td>0.007</td>
<td>3</td>
</tr>
</tbody>
</table>

\[ R^+ = 36 \text{ and } R^- = 0 \]

\[ T = 0 \]

The evaluation shows that T is less than 3 and hence our proposed method performed significantly different from Model-1. For space consideration, the statistical analysis for only N=100 is presented. But for other values of N also, Model-2 significantly outperform the Model-1.
7. Conclusion

We have presented an approach to enhance performance of classifier ensemble for imbalanced data. First concern is to modify random under-sampling approach in order to overcome its limitation of removing necessary instances of the majority class. For this the necessary data is identified first and then random under-sampling technique is applied on the remaining data. Second concern is to design a hybrid approach for constructing the classifier ensemble in which ensembles are formed using two methods i.e. using different training data sets and different learning models. Experimental results of proposed approach are compared with results of classifier ensemble that is constructed using different training data sets only. The evaluation results indicate that the presented approach outperforms the other technique in terms of AUC.

References